PCT/AU98/01023

09/581286

REC'D 12 JAN 1999

WIPO PCT

Patent Office
Canberra

5

I, KIM MARSHALL, MANAGER EXAMINATION SUPPORT AND SALES, hereby certify that the annexed is a true copy of the Provisional specification in connection with Application No. PP 0839 for a patent by CSL LIMITED filed on 10 December 1997.

PRIORITY DOCUMENT

SUBMITTED OR TRANSMITTED IN COMPLIANCE WITH RULE 17.1(a) OR (b)

WITNESS my hand this Fifth day of January 1999

KIM MARSHALL
MANAGER EXAMINATION SUPPORT AND SALES

AUSTRALIA

Patents Act 1990

CSL LIMITED

PROVISIONAL SPECIFICATION

Invention Title:

Porphyromonas gingivalis nucleotides

The invention is described in the following statement:

Porphyromonas gingivalis nucleotides

FIELD OF THE INVENTION

The present invention relates to *P. gingivalis* nucleotide sequences, *P. gingivalis* polypeptides and probes for detection of *P. gingivalis*.

BACKGROUND OF THE INVENTION

Periodontal diseases are bacterial-associated inflammatory diseases of the supporting tissues of the teeth and range from the relatively mild form of gingivitis, the non-specific, reversible inflammation of gingival tissue to the more aggressive forms of periodontitis which are characterised by the destruction of the tooth's supporting structures. Periodontitis is associated with a subgingival infection of a consortium of specific Gram-negative bacteria that leads to the destruction of the periodontium and is a major public health problem. One bacterium that has attracted considerable interest is *P. gingivalis* as the recovery of this microorganism from adult periodontitis lesions can be up to 50% of the subgingival anaerobically cultivable flora, whereas *P. gingivalis* is rarely recovered, and then in low numbers, from healthy sites. A proportional increase in the level of *P. gingivalis* in subgingival plaque has been associated with an increased severity of periodontitis and eradication of the microorganism from the cultivable subgingival microbial population is accompanied by resolution of the disease. The progression of periodontitis lesions in non-human primates has been demonstrated with the subgingival implantation of *P. gingivalis*. These findings in both animals and humans suggest a major role for *P. gingivalis* in the development of adult periodontitis.

P. gingivalis is a black-pigmented, anaerobic, asaccharolytic, proteolytic Gram-negative rod that obtains energy from the metabolism of specific amino acids. The microorganism has an absolute growth requirement for iron, preferentially in the form of haeme or its Fe(III) oxidation product haemin and when grown under conditions of excess haemin is highly virulent in experimental animals. A number of virulence factors have been implicated in the pathogenicity of *P. gingivalis* including the capsule, adhesins, cytotoxins and extracellular hydrolytic enzymes. In

particular, proteases have received a great deal of attention for their ability to degrade a broad range of host proteins including structural proteins and others involved in defence. The proteins that have been shown to be substrates for *P. gingivalis* proteolytic activity include collagen types I and IV, fibronectin, fibrinogen, laminin, complement and plasma clotting cascade proteins, α_1-antitrypsin, α_2-macroglobulin, antichymotrypsin, antithrombin III, antiplasmin, cystatin C, IgG and IgA. The major proteolytic activities associated with this organism have been defined by substrate specificity and are "trypsin-like", that is cleavage on the carboxyl side of arginyl and lysyl residues and collagenolytic although other minor activities have been reported.

P. gingivalis trypsin-like proteolytic activity has been shown to degrade complement, generating biologically active C5a, impair the phagocytic and other functions of neutrophils by modifying surface receptors, and abrogate the clotting potential of fibrinogen prolonging plasma clotting time. The trypsin-like proteolytic activity of *P. gingivalis* also generates Fc fragments from human IgG1 stimulating the release of pro-inflammatory cytokines from mononuclear cells and is associated with vascular disruption and enhanced vascular permeation through the activation of the kallikrein-kinin cascade. *P. gingivalis* spontaneous mutants with reduced trypsin-like activity as well as wild-type cells treated with the trypsin-like protease inhibitor N-*p*-tosyl-L-lysine chloromethyl ketone are avirulent in animal models. Further, it has been shown that *P. gingivalis* grown under controlled, haemin-excess conditions expressed more trypsin-like and less collagenolytic activity and were more virulent in mice relative to cells grown under haemin-limited but otherwise identical conditions. The increased expression of the trypsin-like activity by the more virulent *P. gingivalis* has led to the speculation that the trypsin-like proteolytic activity may be the major determinant for infection or disease.

There has been considerable endeavour to purify and characterise the trypsin-like proteases of *P. gingivalis* from cell-free culture fluids. Chen *et al*, (1992) [J Biol Chem 267:18896-18901] have purified and characterised a 50 kDa arginine-specific, thiol protease from the culture fluid of *P. gingivalis* H66 designated Arg-gingipain. A similar arginine-specific thiol protease has been disclosed in JP 07135973 and the amino acid sequence disclosed in WO 9507286 and in Kirszbaum *et al*, 1995 [Biochem Biophys Res Comm

207:424-431]. Pike *et al* (1994) [J Biol Chem 269:406-411] have characterised a 60 kDa lysine-specific cysteine proteinase from the culture fluid of *P. gingivalis* H66 designated Lys-gingipain and the partial gene sequence for this enzyme was disclosed in WO 9511298 and fully disclosed in WO 9617936.

In order to develop an efficacious and safe vaccine to prevent *P. gingivalis* colonisation it is necessary to identify and produce antigens that are involved in virulence that have utility as immunogens to generate neutralising antibodies. Whilst it is possible to attempt to isolate antigens directly from cultures of *P. gingivalis* this is often difficult. For example as mentioned above, *P. gingivalis* is a strict anaerobe and can be difficult to isolate and grow. It is also known that, for a number of organisms, when cultured *in vitro* that many virulence genes are down regulated and the encoded proteins are no longer expressed. If conventional chemistry techniques were applied to purify vaccine candidates potentially important (protective) molecules may not be identified. With DNA sequencing, as the gene is present (but not transcribed) even when the organism is grown in vitro it can be identified, cloned and produced as a recombinant DNA protein. Similarly, a protective antigen or therapeutic target may be transiently expressed by the organism *in vitro* or produced in low levels making the identification of these molecules extremely difficult by conventional methods.

With serological identification of therapeutic targets one is limited to those responses which are detectable using standard methods such as Western Blotting or ELISA. The limitation here is the both the level of response that is generated by the animal or human and determining whether this response is protective, damaging or irrelevant. No such limitation is present with a sequencing approach to the identification of potential therapeutic or prophylactic targets.

It is also well known that *P. gingivalis* produces a range of broadly active proteases (University of Melbourne International Patent Application No PCT /AU 96/00673, US Patent Nos 5,475,097 and 5,523,390), which make the identification of intact proteins difficult because of their degradation by these proteases.

SUMMARY OF THE INVENTION

The present inventors have attempted to isolate *P. gingivalis* nucleotide sequences which can be used for recombinant production of *P. gingivalis* polypeptides and to develop nucleotide probes specific for *P. gingivalis*. The DNA sequences listed below have been selected from a large number of *P. gingivalis* sequences according to their indicative potential as vaccine candidates. This intuitive step involved comparison of the deduced protein sequence from the *P. gingivalis* DNA sequences to the known protein sequence databases. Some of the characteristics used to select useful vaccine candidates include; the expected cellular location, such as outer membrane proteins or secreted proteins, particular functional activities of similar proteins such as those with an enzymatic or proteolytic activity, proteins involved in essential metabolic pathways that when inactivated or blocked may be deleterious or lethal to the organism, proteins that might be expected to play a role in the pathogenesis of the organism eg. red cell lysis, cell agglutination or cell receptors and proteins which are paralogues to proteins with proven vaccine efficacy. DNA sequences that were considered to be poor vaccine candidates and not selected include those that code for proteins involved in replication, non-essential proteins involved in cellular processes and those proteins present at sites that would be unlikely to be affected by immune mediators such as those found in the bacterial cytoplasm or inner membranes.

In a first aspect the present invention consists in an isolated *P. gingivalis* nucleotide sequence, the nucleotide sequence consisting of or including a sequence selected from the group consisting of SEQ ID NO: 1, SEQ ID NO: 2, SEQ ID NO: 3, SEQ ID NO: 4, SEQ ID NO: 5, SEQ ID NO: 6, SEQ ID NO: 7, SEQ ID NO: 9, SEQ ID NO: 10, SEQ ID NO: 11, SEQ ID NO: 12, SEQ ID NO: 13, SEQ ID NO: 14, SEQ ID NO: 15, SEQ ID NO: 16, SEQ ID NO: 17, SEQ ID NO: 18, SEQ ID NO: 19, SEQ ID NO: 20, SEQ ID NO: 21, SEQ ID NO: 22, SEQ ID NO: 23, SEQ ID NO: 24, SEQ ID NO: 25, SEQ ID NO: 26, SEQ ID NO: 27, SEQ ID NO: 28, SEQ ID NO: 29, and SEQ ID NO: 30.

In a second aspect the present invention consists in an isolated *P. gingivalis* polypeptide, the polypeptide being at least partially encoded by a nucleotide consisting of or including a sequence selected from the group

consisting of SEQ ID NO: 1, SEQ ID NO: 2, SEQ ID NO: 3, SEQ ID NO: 4, SEQ ID NO: 5, SEQ ID NO: 6, SEQ ID NO: 7, SEQ ID NO: 9, SEQ ID NO: 10, SEQ ID NO: 11, SEQ ID NO: 12, SEQ ID NO: 13, SEQ ID NO: 14, SEQ ID NO: 15, SEQ ID NO: 16, SEQ ID NO: 17, SEQ ID NO: 18, SEQ ID NO: 19, SEQ ID NO: 20, SEQ ID NO: 21, SEQ ID NO: 22, SEQ ID NO: 23, SEQ ID NO: 24, SEQ ID NO: 25, SEQ ID NO: 26, SEQ ID NO: 27, SEQ ID NO: 28, SEQ ID NO: 29, and SEQ ID NO: 30.

In a third aspect the present invention consists in a nucleotide probe specific for *P. gingivalis,* the probe including a detectable label and a nucleotide sequence of at least 15(?) nucleotides, the nucleotide sequence being derived from a sequence selected from the group consisting of SEQ ID NO: 1, SEQ ID NO: 2, SEQ ID NO: 3, SEQ ID NO: 4, SEQ ID NO: 5, SEQ ID NO: 6, SEQ ID NO: 7, SEQ ID NO: 9, SEQ ID NO: 10, SEQ ID NO: 11, SEQ ID NO: 12, SEQ ID NO: 13, SEQ ID NO: 14, SEQ ID NO: 15, SEQ ID NO: 16, SEQ ID NO: 17, SEQ ID NO: 18, SEQ ID NO: 19, SEQ ID NO: 20, SEQ ID NO: 21, SEQ ID NO: 22, SEQ ID NO: 23, SEQ ID NO: 24, SEQ ID NO: 25, SEQ ID NO: 26, SEQ ID NO: 27, SEQ ID NO: 28, SEQ ID NO: 29, and SEQ ID NO: 30, or a sequence complementary thereto.

DETAILED DESCRIPTION

Preparation of the *P. gingivalis* library for sequencing.

To determine the DNA sequence of *P. gingivalis* genomic DNA was isolated from *P. gingivalis* strain W50 (ATCC 53978) essentially by the method described by Mamur J. (1961). Cloning of DNA fragments was performed essentially as described by Fleischmann *et al.*, (1995). Briefly, purified genomic DNA from *P. gingivalis* was nebulized to fragment the DNA and was treated with Bal31 nuclease to create blunt ends then run twice on preparative 1% agarose gels. DNA fragments of 1.6-2.0 kb were excised from the gel and the DNA recovered. This DNA was then ligated to the vector pUC18 (*Sma*I digested and dephosphorylated; Pharmacia) and electrophoresed on a 1% agarose preparative gel. The fragment comprising linear vector plus one insert was excised, purified and this process repeated to reduce any vector without insert contamination. The recovered vector plus insert DNA was blunt-ended with T4 DNA polymerase, then a final

ligation to produce circular DNA was performed. Aliquots of Epicurian Coli Electroporation-Competent Cells (Stratagene) were transformed with the library DNA and plated out on SOB agar antibiotic diffusion plates containing X-gal and incubated at 37°C overnight. Colonies with inserts appeared white and those without inserts (vector alone) appeared blue. Plates were stored at 4°C until the white clones were picked and expanded for the extraction of plasmid DNA for sequencing.

DNA sequencing

Plasmid DNA was prepared by picking bacterial colonies into 1.5ml of LB, TB or SOB broth supplemented with 50-100ug/ml Ampicillin in 96 deep well plates. Plasmid DNA was isolated using the QIAprep Spin or QIAprep 96 Turbo miniprep kits (QIAGEN GmbH, Germany). DNA was eluted into a 96 well gridded array and stored at -20C.

Sequencing reactions were performed using ABI PRISM Dye Terminator and ABI PRISM BIGDye Terminator Cycle Sequencing Ready Reaction kits with AmpliTaq DNA polymerase FS (PE Applied Biosystems, Foster City, CA) using the M13 Universal forward and reverse sequencing primers. Sequence reactions were conducted on either a Perkin-Elmer GeneAmp 9700 (PE Applied Biosystems) or Hybaid PCR Express (Hybaid, UK) thermal cyclers. Sequencing reactions were analysed on ABI PRISM 377 DNA sequencers (PE Applied Biosystems).

The sequences obtained are set out below.

DNA sequence analysis

Raw trace data files from the ABI 377 sequencer were manually trimmed using Staden Pregap(Laboratory of Molecular Biology, Medical Research Council, UK) running on a Sun Microsystem computer. Trimmed files were assembled into contigs using Staden Gap v4.1 and exported as a consensus file in FastA format. This consensus was converted into GCG format files and analysed for homology using the BLASTX algorithm [Altschul *et al*] on a non-redundant protein database compiled by ANGIS (Australian Genomic Information Service, University of Sydney). Individual BLAST search results were examined for significant homology by statistical probability and amino acid alignments.

The results are set out in Table 1.

It will be appreciated by persons skilled in the art that numerous variations and/or modifications may be made to the invention as shown in the specific embodiments without departing from the spirit or scope of the invention as broadly described. The present embodiments are, therefore, to be considered in all respects as illustrative and not restrictive.

Dated this tenth day of December 1997

CSL LIMITED
Patent Attorneys for the Applicant:
F.B. RICE & CO.

References.

Mamur, J. (1961) A procedure for the isolation of deoxyribonucleic acid from micro-organisms. J. Mol. Biol. 3, 208-218.

Fleishmann, R.D. et al. (1995) Whole genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269, 496-512.

Altschul SF, Gish W, Miller W and EW Myers. (1990). Basic local alignment search tool. J. Mol. Biol. 215:403-410.

SEQUENCES

Seq ID # 1
Length: 389

```
   1  ttcgtccaca tcgtcgcctt cggggatcac cgtgatctgg tcacaccggg
  51  ctgacaggaa aggctccttc tcgcagagct tcttaccggt cagcacattg
 101  ccatcgatca cacgctcgtg agattccttt attgtcagtc ggccggggaa
 151  ggaagcaaag acattgcaac ccggcataat acggacgtat ccgtgagctg
 201  cagcgtctga gccggtcatg gcaatcattc tggtaaaatc ggctttgccc
 251  gtaagcagga aacgtccgat cacgatcagg tcggtagcct tgagcgtcca
 301  caccgtttcg ccccgattga ttggcttccg tatgattgat cagcacgccc
 351  acttttacct gccggatgaa gtcccgtgtt acttcctac
```

Seq ID # 2
Length: 912

```
   1  aacgattgtc ggctgattct tgcttcctgc acgatgcagg acncgattgt
  51  cagttgattc ttgctccctg cacgatgcag gacgcgattg tcggctgttt
 101  cttgctccct gcatgatgca ggacgcgatt gtcagctgat tcttgctccc
 151  tgcacgatgc aggacgcgat tgtcagttga ttcttgcttc ctgcacgatg
 201  caggacgcga ttgtcagttg attcttgctt cctgcacgat gcaggacgcg
 251  attgtcaact gattcttgct tcctgcacga tgcaggacgc gattgtcagc
 301  tgattcttgc ttcctgcacc gatgcaggac gcgattgtca gctgattctg
 351  ctcccatcaa tgcgctaact atcaagctgt ttgcaactat tttataggac
 401  tttcattgaa gtcttttgcc gcagagctga ttcttaagtg tttttcagat
 451  tacttgaggt ttgcagagag atcgcatgaa gctctccttt cttcgtcaaa
 501  tcaatgcttg tgtctgtctt gatcaatatg agaggggttt attgtgcaac
 551  ggtctcaagc tgtaaaaccg gcagctgtgt atagaaacag tctttcggtg
 601  atatggcaat cgaacttcct aactgcccaa attttaccgg acagcaataa
 651  ctgattatat ggggttagtc catcgggcgg actttctctt cgacaaaggc
 701  gatttcgctt tcgtttaagc cgtatttgcg gtagagttgg cgatcaattt
 751  ccgccaccgg ctgtgtccaa tcgatgtccg attctgctgt gaagtcttgc
 801  agggggacta atcgccaagt ttctttgggg ttatcttgcg ttgccttgag
 851  gatacccagc atcgtgcgag caaacttcgt cttcacatag cgcagacaag
 901  cctctgcctc gg
```

Seq ID # 3
Length: 408

```
   1  gagaagaaag ctcctgcact gaggaaagga gcgttaggct tgtgagtaat
  51  ctcggacaga cgctcattca cggctgtagt gatcacctgt ttcatatagt
 101  cttccacaag tccgaatatc gatcctcgca cttcttgagg agtggggtcg
 151  ctcttgaagc tgatggagag ctgcgtggta gtagcctcag catcggtagc
 201  aatggctacg ataggctcat cgttgtcctc taccggcgta tagatacgct
 251  ctgctggatt cacgggagca ggaacgtcct tgaagagttc tttgatcttg
 301  ttctccacat agtccacatc gatatctccc acgatcacca gaccttgcag
 351  gtcgggacga taccatttct tataatagtt gcgcagctca tcatgcttgg
 401  aagttgac
```

Seq ID # 4
Length: 643

```
   1  cgtgtgagca acactttcct tggggccatg cagacccaga gcacttgtcc
  51  cacttgccac ggagagggtg agatcatcac gaagccatgc tccaagtgta
 101  agggcgaagg tgtggagatc ggcgaagagg tgatctcatt ccacatccct
 151  gccggtgtag ccgaaggaat gcaaatgtcc gtgaacggca agggaaatgc
 201  cgcgccccga ggaggcgtga atggcgactt gatagtcgtg atcgccgagg
 251  aaccggatcc gaatctgatc cgcaatggca acgatctgat atacaatctg
 301  cttatatccg ttccgttggc tataaaagga ggtagtgtgg aagtgccgac
 351  gatagacgga cgagccaaga tccgcatcga ggcggggaca caacccggca
 401  agatgctgcg tttgcgcaat aaggggggttg ccagcgtaaa cggctatggc
 451  atgggagacc aactggtgaa tgtcaatgnc tatatccccg aatcgatcga
 501  tgccnaagat gagcaggcta tcgcagcgat ggaaaactcg gacagcttca
 551  aacctaccga tgctgctcgt aggatatnga caagaaatca gagagatgct
 601  ggattgaaag acaatgccat tgtatggtac ttgacctgag aaa
```

Seq ID # 5
Length: 311

```
   1  gggcggcgag ccggtttgga atacggccgc acgccaaggc atccgtaccg
  51  gtgtctactt ttgggtagga tccgaaacgg ntgtgaacgg aaatcggccg
 101  tggcggtgga aaaaattctc ctccaccgtt ccgtttcgtg accgtgccga
 151  ctccgtcatc gcgtggctcg gactgcccga aaaggagcga ccgcgcttgc
 201  tcatgtggta catcgaggag ccggatatga tcggacacag ccaaactccc
 251  gaaagcccgc tgacactggc aatggtacac cggttggaca gtgtggtcgg
 301  ctatatccgc a
```

Seq ID # 6
Length: 366

```
   1  gccgtgtaag cgcaataggg tcaagcgttt ggtcagggag gcttatcggc
  51  tcaacaaaca cctcctgaac gatgtcctcc aagagagaca gatctatgct
 101  actattgcat ttatggtagt atcggatgaa cttcctgact ttcgtacagt
 151  ggagagagcg atgcaaaaga gtctgatcng aattgccgga aatgtacctt
 201  catcggcttt gaaaaacgag tanatacgat gcgactgatc aaggcttttc
 251  tcgtgcaact cttactgctc cccatttttct tctacaagcg gtttatatcg
 301  ccgcttacac cgccttcatg ccggtttacc ccctcatgtt cgtcctatgc
 351  catccgaagc cttacg
```

Seq ID # 7
Length: 482

```
   1  ctacttatct ataaactcga atcttacgaa ctgttccgca agatggtaga
  51  agccatgaac cgtaagaccg tagcgatcct aatgcgtgct cggataccgg
 101  taccggaggc tccttcccaa gaagagctgg aacacaggcg gcaaatagaa
 151  atccgacatg cagccgaaca acgtacggac atgagtaagt atcggacaca
 201  aaaagacgat atagaagccc agcagaaagc acaaagggat gcggcaagca
 251  gacctcaggg tgcagctgct ccccagacac cgataagaaa cgagaataag
 301  atcgggcgaa acgatccttg tccttgcggt agtggcaaaa agttcaaaca
 351  gtgccacggg cgtaacctgt aaaaagattt atgagagaat caccgactat
 401  ggtatagaat agtctgngat tctcttttta ttttttctct ctacccgcat
 451  ataaaaaaga ctatgatcct atctatgacc gg
```

Seq ID # 8
Length: 500

```
   1  cggcgatggg cgatgttgcc ggaatggcct atcttgattc catgtcgaat
  51  gagaaggtct ggttcggcta cacgctgaaa gaagctcaag cccagcaaat
 101  tggtcttggc cttgacttaa agggggggtat gaacgttatc ttgaaactta
 151  acgcaagcga tctgcttcgt aacctctcta acaaaagttt ggatcccaac
 201  ttcaacaaag ctctggagaa tgctgccaag agcacggagc aatccgactt
 251  catcgatatt ttcgtgaagg aatatcgcaa gctcgatccc aacggtcgct
 301  tggccgttat cttcggttcc gggtgacctt cgcgaccaga ttaccgcaaa
 351  gtctacggat gcagacgtag ntgcgtctgc tcaaagaaaa atataatagt
 401  gctgtagaag cttcgtcaat gtgctccgtg gctcgtatcg atgctttcgg
 451  tgtggntgca cctaatttgc agcgattgga aggacagggg cgtatncttg
```

Seq ID # 9
Length: 352

```
   1  aagccgaaca tgcgtggtac gntgaccagc gtgtagcata gtaccggcac
  51  gtagatggtc agataggcca gccccttgcc aaagacgata cggagtgcgc
 101  cggagtagtg gctgatatcg gtaggtacga ggtcgtgaaa cttattcctc
 151  tcacgcgccg taccggccga gagtcctatg ccgagcagga gcgtctgctg
 201  tatgatcagg atgagtacgg caggtagcag gaaggaagcg aaaccgacgg
 251  tcgggttata cagtgccacg tcttcatagt cgatggggta aganatgatc
 301  tctccctgac gctcggtagc tcctacgctg cgcgctatct tgatctcttt
 351  gt
```

Seq ID # 10
Length: 516

```
   1  agcgatcaaa gcggcagtgg aactgaccga tcgctatgta tccgatcgtt
  51  tcttcccaga taaggcgata gatgccatgg acgaggccgg cgcgagcgtc
 101  catatcacca atgtggtggc tccgaaagaa atcgagatac tggaggccga
 151  attggcatcg gtgcgagaga acaagctctc ggccgtaaag gctcagaact
 201  acgaactggc tgcctccttc cgcgatcagg agcggcgcac tcagcagcag
 251  atagcggaag agaagaaaaa atgggaagag cagatgtcca agcaccgcga
 301  gacggtggac gagaatgtag tggcgcatgt agtgggcgtt gatgacaggc
 351  gttccggctg agcggctgag cacgggcgaa ggcgaacgtc tgcgcacgat
 401  ggcagatgat ctcaagacca aagtagtagg tcaggacaca gccatcgaaa
 451  ggatggtgca tgccatccag cgcaatcgtc tggggacttc gcaatgaaaa
 501  gacccgaacg ggtctt
```

Seq ID # 11
Length: 401

```
   1  ttcgaggcat tcctgcgcta tgggctcaag cctacattct tgactcctcc
  51  atccatgcag cgcgctgtcg agatgttcga ctaccgctca ggagaaaaat
 101  acgaatggaa tgcttacccc acctatgaag cctatatcag catgatggaa
 151  gagttccaaa caaagtatcc atcactttgt actacttccg tcattggcaa
 201  gtccgtaaag gatcgtaaac tgatgatttg caagctgacg tcctctgcca
 251  atacagggaa aaagcctcgc gtgctctata cttctacgat gcacggagac
 301  gaaacgaccg gatatgtggt nctgctccga ctcatagaac atctgctgtc
 351  gaactacgaa tccgatccga ggattaagaa cattctggat aaaacggaag
 401  t
```

Seq ID # 12
Length: 553

```
   1  ggcttacacc ctgattttcc tgtgcacgaa aagttttctt atacgtttcc
  51  tccaagaccg gttcgccgga tccttttttc ttttttattt cgaacggact
 101  attctgccag ccatccgtat ggtagaagaa atagttggta ttgctactga
 151  acttaccgat gtttatatcg agcgaagtat tggtttgccg atcgttgnac
 201  ttcgatatgc gatgtgacgt atatgcactc agtcgattcg tattcttttt
 251  tgtgatcaca ttgattaccc cggcgatggc atcggatccg tagagcgaac
 301  tcgaagcacc tttaccagtt cgatccgttc gatctgatca ggagaantac
 351  gactcaaatc ggcctgaccg cctacatcgc cgtacacacg cttaccatcg
 401  ataaggatga ggatatactt actgctaagg ccgntcagct gcatgaaaga
 451  gcccatcaga ttggggccga agtcaaaaga cggactcagc cctgcataag
 501  gcctcggaag taggagccga gaaagaggct atgtccttag cggtaaggac
 551  ttc
```

Seq ID # 13
Length: 450

```
   1  cacacaataa cttgtgctgc attttcttat tgtattttca gaaaaagaag
  51  caaccccaat gacactccac tttgtgtcgt tggggttgtt gctttagaag
 101  taagtgatca gattatttct tagctgcttt gttgagcttg cgcttttgga
 151  tctcgtaggc aaggggtgta gcaacaaaga gcgtagagta tgtaccgata
 201  acgataccga gcaggatcga gaacgtgaaa ctacgcatcg tagcacctcc
 251  aaagatgaag attaccaaca taacgataaa cgtagtcaaa gacgtattta
 301  atgttcgacc caatgttgaa ttaagggcat cgttgatcac ctgatagcga
 351  tctctgttgg ggtacaattt catcgtctct cggatacggt caaatacaac
 401  cacggtgtca ttgagcgagt naccgatgat agccagaata gcagcgatga
```

Seq ID # 14
Length: 383

```
   1  ggaatgtgcc gaacgaacct ccgctcaaat cgacacggcc ataaggagcc
  51  agtcccaaat tatccgtacg catattgaca cttgccccaa aagctccggc
 101  accattggtg gaagtaccca cacctcgctg cacctgaagg tcttcgatgg
 151  aagaggcgaa gtcgggcata ttcacccaaa agacggactg agattcggag
 201  tcgttgaggg gtactccatt ggtagttatg ttgatgcgat tggcatcggt
 251  gccacgcacg cgaaagccgg aatatccgat acccgtaccg gcatcgctgg
 301  tggctaccac ggagggagtc agcatcagca gataggggat gncacgacca
 351  taattggact tggaaaggtc ggccttgcga acg
```

Seq ID # 15
Length: 477

```
   1  tcggagagag acgtttttcc ttcgaaaaga taactgccat cccccaaaac
  51  cttaaagggg agttcttcct catcgtactc gtccgtaatc tcgccgacga
 101  tctcttccca atatgtcctc cattgtgatc agtccgcaag tgccaccgaa
 151  ctcatccaca acgatggaga catgcacctt attggctctg aactcctcga
 201  gcaaatcatc tatgcgcttg ttttcgggga caaaatatgc tttacgaatc
 251  agaggatgcc agtcgaattc atcgccttta tccatgtgtg ggattagatc
 301  tttgatgtaa atcacccctt tgatattgtc ttctgacccc tctgaaacgg
 351  gaagtctgga ataacccgac gaaacaacga agtcaagcat cttacgaaat
 401  ggccagctca gatccacatc cacaatatcg atacgcggga accatggatt
 451  tcgcaggctg gcttattata ggaattg
```

Seq ID # 16
Length: 486

```
   1  gctcattttt acctttcttc gtttgaaatg aaaacgactc cgtttgcagc
  51  acgagctcca taaatagatg ttgcagaagc atctttcaaa acggacatag
 101  attcaaaatc attcggattc atcgtagcca caacatccaa agaagtttgc
 151  ataccatcca cgatatacaa tggtgcagag cttgccccca acgaccctgt
 201  accatggatc tccacagaag cgacggcagt agggtcaccg gatgtagtca
 251  taacctgcat accggctacc tgaccttgga gggcatccat gatattggca
 301  acgggctttt ccgcgagctt ttcgctggac actttggcca cagaaccgga
 351  aacagtgctg agtttctgtc ccgtaccgta acccaataca actacctgct
 401  ccagaacctt agagtccgga tccagtacga tcttccatcc acatttagcg
 451  aatggcgaac tcctttggta atcatacccg ggaaat
```

Seq ID # 17
Length: 386

```
   1  ccgaacatct catcacacnc aatagggaag acctcagtgg catagccata
  51  gccgtagcga tggagggcat tcgcccgata ctcatcgaag cgcangcttt
 101  ggtcagctcg gccatttatg ccaatccgca gcgttcggcc acgggcttcg
 151  atattcggcg gatgaacatg ctcttagccg tactggagaa acgtgccggc
 201  ttcaagctca tacagaanga tgtgtttctg aacattgccg gaggtatcaa
 251  aatagccgat ccggctacgg atctggccgt tatctcggca gtgctggcgt
 301  ccagtctgga catcgttatc ccgccggccg tatgcatgac gggcgaagtc
 351  ggactctccg gananatacg tcccgtgagc cgcatc
```

Seq ID # 18
Length: 1013

```
    1  gattatgatg aagagacttg ggggaaatgg tgtgcacagg ccgatgccga
   51  cacactggca ggagctttgt ctttcttcct ccatgcagcg aacaagggga
  101  tcgaggtctt ttacgtcacc aaccgcagag acaatntgcg cgaagcaact
  151  nttcagaacc ttcagcgtta cggattcccc tttgccgatg aagaacattt
  201  gcttacgacc catgggccat ccgacaaaga accccgtcgg ctcaaaatac
  251  aagaacagta tgaaatagta ttgctcatag gagacaactt gggcgacttc
  301  caccacttct tcaatacgaa agaagagtcc ggacgcaaac aggctctggg
  351  cctgacagcc ggggagtttg gccggcactt catcatgctg cccaatccca
  401  actacggatc ttgggaaccg gcatggtacg gcgggaagta tccgccactg
  451  cccgaaagag acaaagcact taaacaactg cactcacaga acagcagata
  501  gctccttaag caaacacatc gaatagacag actcacacta tggacaacaa
  551  acgactaagc aaaatagaaa gactgctcca gaaagaactc agcgagatat
  601  tcctgcggga tgcgaaatcc ctgccgggcg taatagtttc ggtaacgaac
  651  gtacgtgtaa gtcccgacct cagcatcgca cgtatacacc tgagtatatt
  701  cccatccgag aagagcagcg agattcttga gagcatcaaa cacaatacaa
  751  agacgatccg ttatgacctc gggcagcaag ttcgtaccca actgcgcaag
  801  ataccggatt tgacattcta catagatgac tctctggatt atctggagaa
  851  tatagaccgt ttgctcaatc aataagaaac ggtcgctctc tatcaagacg
  901  ctgtgaactt ccctttttttc atagcccgcc gttacttgtt ctcccgcaaa
  951  agattcagtg cggtcaatgt ggtttcgctc gtttcagcga tagctgtctg
 1001  cgtggtctct tcg
```

Seq ID # 19
Length: 445

```
   1  aacaactaat gtctcacaaa ttaatttaag aacagagatg aaaaaactga
  51  ttttagcgac tttgggactt atggccattg ccatgctctc atgttcaagc
 101  aacaacaagg atttggagaa caaaggggag gctactcttt tggtaacgtt
 151  tggtagctcc tataaagctc cacgcgaaac ctatgcgaag attgagaaga
 201  cttttgccgc agcttatccc gatcaaagga taagctggac atacacgtct
 251  tctattatcc gaaagaaact ggctcagcag ggtatttata tcgatgctcc
 301  ggatgaggct ttggagaaat tggctcgtct gggttataag aagatcaatg
 351  ttacagagtc ttcatgtgat tcccggccga gaatatgatg agatgatcga
 401  ctttgttcaa taattttaag gcagcacata gtgatattac tgtga
```

Seq ID # 20
Length: 488

```
   1  cggccgaagc ccagaccgat caatgtctgt tcgatcgcag catgatagtt
  51  gccctgctgc atgagagaga gggtctcgct catattcgta taatgctcta
 101  tcagtcggat atagtcatcc gattcgtagt ccgtacgtcc ggccatttca
 151  tcggacagac gccgtatctc ttcctctatt tggcgaatat cgttgaaagc
 201  ctgctcgacc tcttcgtaaa ccgtgtgtcc gtcctgcaaa cgcatcacct
 251  gcggcagata gcctatgcgg atccccttgg ggcgtgctat gtgtccggat
 301  gtcggttctt ccatgccggc aatcagcttg agcagcgtac tcttgccggc
 351  accgttcttc cctacaagag cgatacggtc gcgcctgttg atgacgaatg
 401  atacctgatc gaagagcaga cgggtgccga aatcgacagt caggttattg
 451  acggagatca tgacttcgtg ctcattcgnt tgatgatg
```

Seq ID # 21
Length: 836

```
   1  cgcattccgt cggatatgct catcggcaaa ctggaatcgc tcatcgcttc
  51  gtacataacc ggatcgatcg gaagagaaat agcatgaaga aggaggtgtg
 101  tcaataatca tggcgcacct ccttgcatta atatgggacg gtcgggtaca
 151  ccggctattc cgagactcct taaggagtcc ctaccgagac ccttaaggag
 201  tctccaccaa gacccctaag gagtctccac cgagatccct aaggagtccc
 251  taccaagacc cctaaggggt cccaacagag actccttagg ggttcctcaa
 301  tgctttactt caggaggggt tcgtgcggtc ttataatcta ttcgaatgga
 351  gacatcggga gcagtgaccg gcgaaaggaa gccgaagctt agcgaatctt
 401  accgtcgaac agattgatga tgcggccggc actacgtgca tcgtgctcgg
 451  agtgcgtcac catgacgatg gttgcacctt cgcgattgag acctctgagc
 501  agttccatga catcggctcc gtttttggag tcgaggttac ccgtgggttc
 551  atcggcgagg atgagcttcg gattggccac cacggcncgg gcgatagcca
 601  cgcgctgctg ttgtcctccg gagagctgat tggggaagtg gccggcccgg
 651  tggctgatgc tcatcttgcg cagtgcctcc tccactcgct ctttccgctc
 701  ggaagccttc acacccagat ngacgagcgg caactccacg ttctcgctta
 751  ccgtcatctc ttcgatgang ttgaagctct ggaatacgaa gccgatattg
 801  cccttacgga cgcagtcctg tctttttccc ggaagt
```

Seq ID # 22
Length: 365

```
   1  cggcaaagag atattgaaag gaatcaatct ggagatcaat gccggagaga
  51  ttcatgctat catggggccg aacggatcgg ggaaaagtac gctctcttcc
 101  gttttggtgg gacatccctc ctttgaagtc acggatggag aggtgacatt
 151  caatggaatc gacctgctcg aactcgaacc ggaagaacgt gcacacctcg
 201  gactctttct cagtttccaa tatccggtcg agatcccggg cgtcagcatg
 251  gtgaatttca tgagggcagc tgtcaatgaa cataggaaag cgatcggagc
 301  agaacccgta tnggcaagcg acttcctcaa gatgatgcga gagaagcgtg
 351  ccattgtgga gctgg
```

Seq ID # 23
Length: 640

```
   1  ccactttaac tataaagcct ctatactttt atagtataaa gcctgcgagc
  51  tttatagtcg gaagtattaa agggatgatt gtcgtgctac acttgtcaag
 101  aaaaaggatc agaacggata gcctactgca atgcgccaag cgaaattgga
 151  agaaaggttt gggcgtgtga tagcccattt gtaacgccct gtctgctgag
 201  gatcgtaggc tttcagtccg gcatccagcc gcacaaggaa ataatcgaag
 251  tcgagacgaa gccccagacc gtaggccaaa gctatttcct tgtagaagcg
 301  atcgaaacga aagagaccgt cctcctgatt ctcatactcc tttatcgtcc
 351  agacattgcc ggcatcgaca aaagctgctg cgcgaaactt ccagaacagc
 401  tttgtcctgt attcgacatt cagatccaga cgaatatcac ccatctgatc
 451  gaagaaggtc ttgtccggag tcatcttcat actccccggg ccgagggtac
 501  ggacactcca gccgcgaacg ctgttcgatc ctccggcaaa gtaacgtaac
 551  taaagggtat atggcgagca ttgccataag ggaaagccag tccgaaaccc
 601  agattgcagt gccaaagtat tggccttttc gagagaacgg
```

Seq ID # 24
Length: 771

```
   1  ccaggacaat gcaaattatt tccatcgtct gcgagaaatt acccttgaaa
  51  tcagcaacac gaagttggtg ccggcctctc aacttccaaa gtattggaat
 101  ctgaacaaag aatctctgct tgctctgatc gaagaatcct tatacggcat
 151  ccatggtaca gtgacttccg ctgcgaacgg acagcctctc aaatgccaga
 201  tcttgataga aaaccatgac aagcgcaact ccgatgttta ctccgatgct
 251  accacaggct actacgtacg tcctatcaaa gccggcactt atacggtgaa
 301  atacaaagcc gagggttatc ctgaggcaac tccgnaccat taccgatcaa
 351  ggacaaagaa accgtcatca tggacatttg cattgggcaa cttcggttcc
 401  tctgcctgta cccgatttca cagcttctcc tatgaccatc tcagtaggcg
 451  aaaggcgtcc aattccaagg atcaaacgac aaataacccc acgaattggg
 501  agtggacgtt cgaaggcgga cagcctgcca tgagtacaga gcagaatccg
 551  ctcgtatcct atagtcatcc cggtcagtac gacgttacgc tcaaagtgtg
 601  gaatgcaagt ggttccaaca cgattacgaa agaaaaattc atcactgtca
 651  atgccgatat gcctgtagct gaattcgtcg gtaccccgac ggaaatagaa
 701  gagggccaga cggnatcttt ccaaaaccaa tccaccaatg ccaccaacta
 751  cgtatggata ttcgatggcg g
```

Seq ID # 25
Length: 521

```
   1  gcattgattg taaacagctt cttcacattg ggcgtatngg cttctttnca
  51  tgccgtgctg accctctcgg gtatngcagg tttggtgctg acgctgggta
 101  tggctgtgga tgccaacgta cttatcttcg agcgtatcaa agaagagctt
 151  cgtgccggta agactccgat tcgtgccgtt acggatggtt atggcaacgc
 201  tttctctgcc atcttcgact cgaacgttac gactattatt accggtatca
 251  tcctattcct ctacgggacg gggccgattc gcggttttgc cactacgttg
 301  attatcggtc ttatcgcttc tttcattacg gctgtcttct tgactcgtat
 351  cgtcttcgag aaactggcga aaaaaggtcg tttggataag attacattca
 401  ctacgagcat tactcgcaat ctccttgtca atccctcata caacatcttg
 451  ggtaagcgca agaccggctt tatcattccc ggtgattatc atcgtttggg
 501  acttatagct tcatttacaa t
```

Seq ID # 26
Length: 594

```
   1  cgactcccga tgttccgata atagagatgc cgttccgtaa gccgagtcgg
  51  ggattgaatg tctgggtaca gcctctcggc cttcgggtac gctaatcgta
 101  atatccacac ctccctgcgc atagagtcgg cgtacctctg ctgtcatcat
 151  tcgtcgaggt acgaggttga tagccggacc tccgacctcc agaccgaggc
 201  cggggagcgt cactaccccc acccccttcac cctgcaggaa gcggacttcc
 251  tcatgttcgg gattgagcct gatcgtagcg cataccgcca tgccattggt
 301  cacatccgga tcatcacctg catctttcag gactgcggat acgacagcat
 351  cttcttcctc tcgaatttcc gctatgggca gactgactat ttcgcccgaa
 401  ggcaattcta cgggagcttc ggcgagagag ccaagcccca tcaatcggta
 451  catggctgct actactgcag cggtagctgt ggtgccggta gtgaatccgt
 501  tcggagtgaa aagaatcccg gacgaagcgt ctaccggccg tctaaaccga
 551  caggccgcta cggaatgaag aagaagcaaa ggggacgtcc acgg
```

Seq ID # 27
Length: 587

```
   1  ccctgcattg ataaccactt tacgcttgtt gatagcagtt accctaccca
  51  ttgccacttg cttttcgtgg atggtattga gcgtttngtc ataggcagct
 101  tcctgctcct gacgactcac ggcagcatgg acgccaccgg cctcaaactc
 151  ttcccagttg aaatcctctc tgggctgaat gttctttaag ttttccattt
 201  gttttttaat cgtcttgttg ctaatagatt ggacattata tggtcatatc
 251  tctcaaaagc acggnaaagt tacaataact tttcgttact ctcttcattg
 301  tgatcaacct gcagattccc cccaccggtg caacattaag caagcgacct
 351  caggattgac tcccaagagt caaccgaaat caggaaacga cactatttga
 401  aattacaatg ttgcaatatc gatcttggcg taaaactgat cggaaacggt
 451  cccgatgttt cttcacaatt actgcttttt ttgacctcct caagcctcat
 501  tttttcagta cacgtcacgt cagtcgtcag tataaaaaag tgacgcgtgc
 551  ctttnctgaa aggcgcgcga gaatttcccg tttgcgc
```

Seq ID # 28
Length: 740

```
   1  gtatcggaag gccgcaaccg caccaaggca cagatcgaca gcatcgctca
  51  aggccgtgta tggctcggcg acaaagctct tgcactcggt ttggtggatg
 101  agcttggagg tttggacaca gctatcaaac gggccgcgaa gctggctcag
 151  ctcggtggca actacagcat agagtatggc aagaccaagc gcaacttctt
 201  cgaagagttg ctctcctcat cagcagcgga tatgaagtct gccatcctga
 251  gtaccattct ctccgatccg gaaatagaag ttctgcgcga actccgctcc
 301  atgccgcccc gtccttcggg catacaggca cgtctcccct attacttcat
 351  gccgtactga taaatgagac aaccgtaatt gctgaagaga tggatgcgcc
 401  ccgtatcaac aagtggctca aaccgctttc cgccctctac ggggtgggcg
 451  tgaggttgcg caactacctc ttcgacaaga acgtcctgat ttcgaactct
 501  ttcgacatcc ctatcgtctg tgtaggcaat atcaccatcg gcggcaccgg
 551  taagacaccc cacgtagaat acctgattcg gctcctgcat ccacgctatc
 601  gtgtagcagt ggttagccgc ggctataagc ggaaaaccaa agggatgatc
 651  gttgcaaccg aaggatcgac tgcatgggat ataggagacg aacctcgtca
 701  gatcaaacga aaatatccgg acctgaccgt catcgtggat
```

Seq ID # 29
Length: 613

```
   1  ctcccgattc gcccataagg tgagcctgca caacagttgc cacggtgtgc
  51  gcgaactgca tctgtccacc cccagtgaag tgcaccgacc gtaccacaac
 101  aaggtgcgcc ggctattgga gatggtgcag ggcatagagg tattcgagcc
 151  gaagcgaata gacgaatgct gcggtttcgg cggtatgtac tcggtggagg
 201  agccggaggt atccacctgt atggggcatg acaaggtgct ggatcacata
 251  tccacaggtg cggagtacat cacagggccg gacagctcgt gcctcatgca
 301  tatgcaggga gtgatagaca gagagaaatt gcccgatcaa gacaattcat
 351  gcagtagaaa ttttagcagc aaacttattg agtacgaagc atagcgaagc
 401  ggctgcccgc tttttggaga ataagtccgg agcccaagtg gcatgacgag
 451  acgctctgga atggtgcgcc acaaacgcga catccagcgt gatacggtgc
 501  cccgagtggg gaaagatctg cgccaactgg gctcatgaaa tcaaacgctt
 551  caatgtgaca cacttggatt ganctgctgc tgcgatttga agaaatgctt
 601  cgtcgaaccg gtg
```

Seq ID # 30
Length: 560

```
   1  tgggtatagc cagagctttg ctggcgaagc ctgcgttgat cctggccgac
  51  gaacccacag gcaacctcga ttcggtgacc ggattgcaga tcgcttctct
 101  gctctacgaa atcagtaagc agggcactgc agtacttatg agcacgcaca
 151  acagcagcct gctgtcgcat ctgccggcac ggacattggc cgttcgtaag
 201  aatggcgatg cctcctcttt ggtcgagctt gagtgcagat gctgtttcaa
 251  gaaaaaatac ggaaatagat tagcacgata agatcaggaa ttgaaagttc
 301  tcaaatttgg cggtacgtct gtaggagatg ctgaagcgta tccgcaagtg
 351  ttgcccgact gattactttc ggtaaaagga agaaaaatta tagtcctttc
 401  ggctatggcc ggaacgacca attcgcttgt cgaaatagcc tcacaccttg
 451  tcaaacgcca atgtggcaca ggcaaagagg gtgtgccaag gtgttgcgag
 501  agaaatatca tcgcgaaata aatgctctat ccaaacgtnc ggataccttg
 551  agcgcagcca
```

Table 1

Seq ID#	Description	Accession number	% identity	overlap (aa)
1	48kD outer membrane protein of Actinobacillus pleuropneumoniae	Q44130	29.4	126
2	Eukaryote outer membrane protein, TCR junction sequence	E259352	40.7	113
3	Periplasmic zinc metalloprotease belonging to the insulinase family, Escherichia coli	P37648	22.2	126
4	Heat-shock protein DNAJ of Legionella pneumophila	P50025	42.2	173
5	Eukaryote plasma cell membrane glycoprotein (alkaline phosphodiesterase I)	P22413	42.6	101
6	Alpha-hemolysin gene, hlyA, Aeromonas hydrophila	L36462	55.6	36
7	Protein translocase SecA subunit, Escherichia coli	P10408, P75642	40.7	118
8	Protein-export membrane protein (secD), Helicobacter pylori	AE000652_12	41.2	34
9	Hypothetical integral membrane protien, Helicobacter pylori	AE000647_20	29.2	65
10	ATP-dependent CLPC protease, Mycobacterium leprae	P24428	40.1	172
11	Zinc-carboxypeptidase precursor, Streptomyces capreolus	P39041	30.3	99
12	Hem B receptor, Porphyromonas gingivalis	U87395	37.6	109
13	Protein export protein, Helicobacter pylori	AE000652_11	45.8	96
14	Haemoglobin receptor, Neisseria gonorrhoeae	P72073	28.7	115
15	Haemolysin, Helicobacter pylori	AE000647_24	37.7	146
16	Outer membrane protein, Bacteroides thetaiotaomicron	Q45780	39.9	168
17	ATP-dependent protease, Helicobacter pylori	AE000542_5	40.7	123
18	Acid phophatase precursor, Flavobacterium meningosepticum	O08351	41.1	129
19	CBIK protein involved in Cobalmin biosynthesis, Salmonella typhimurium	Q05592	42.3	97
20	ABC Transporter, Haemophilus influenzae	O05519	34	153
21	ABC transporter, Bacillus subtlis	AF008220_56	49.3	240
22	ABC transporter, Mycobacterium leprae	E343546	56.3	87
23	Cysteine protease, Trypanosoma brucei	S12099	33.6	107

Table 1 (cont.)

Seq ID#	Description	Accession number	% identity	overlap (aa)
24	Suface antigen gene, Methanosarcina mareii	X84710	42.9	126
25	Protein-export membrane protein SECD, Haemophilus influenzae	P44591	51.5	132
26	Hypothetical protein involved in Cobalamin synthesis, Methanococcus jannaschii	Q60342	37.5	168
27	Haem uptake protein B , Bacteroides fragilis	Q45140	53	66
28	Virulence-associated ABC transporter, Francisella novicida	Q47909	36.4	99
29	Hypothetical secreted protein, Helicobacter pylori	AE000535_9	28.4	102
30	ABC transporter FTSE, Escherichia coli	P10115	48.4	64